IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, p-ISSN: 2278-8719, www.iosrjen.org
Volume 2, Issue 9 (September 2012), PP 24-42

In the Presence of Reverberation and Interfering Noise Sources, Joint Localization and Recognition of Speakers, With the Binaural Scene Analyzer

Vandana Gupta
ABES Engg.College Ghaziabad

Abstract*—In this paper, we study about a binaural scene analyzer that is able to simultaneously localize, detect and identify a predefined number of target speakers in the presence of spatially positioned noise sources and reverberation. Compared to others, this proposed system does not require a priori knowledge about the azimuth position of the target speakers. The proposed system contains three main components: binaural localization, speech source detection, and automatic speaker identification. First, a binaural front-end is used to robustly localize relevant sound source activity. Second, a speech detection module based on missing data classification is employed to determine whether detected sound source activity corresponds to a speaker or to an interfering noise source using a binary mask that is depend on spatial evidence supplied by the binaural front-end. Third, a second missing data classifier is used to recognize the speaker identities of all detected speech sources. The proposed system is systematically evaluated in simulated adverse acoustic scenarios. Compared to state-of-the art MFCC recognizers, the proposed model achieves significant speaker recognition accuracy improvements.*

Keywords—*Automatic speaker recognition, binaural processing, computational auditory scene analysis (CASA), mask estimation, missing data.*

I. INTRODUCTION

While being constantly surrounded by a variety of different acoustic sources, among them concurrent speakers and environmental noise, the human auditory system is capable of recognizing a single target speaker and selectively following a conversation [1], [2]. According to Bregman [3], the underlying perceptual mechanisms that enable the human auditory system to perform *auditory scene analysis* (ASA) can be divided into two stages: First, the acoustic input is decomposed into a number of segments. In a second step, individual segments that are believed to belong to the same acoustic object are grouped to form a coherent stream.

The noticeable capabilities of the human auditory system to process an arbitrary target source in complex acoustic scenes have inspired a new field of research, named *computational auditory scene analysis* (CASA), that attempts to achieve human performance with computational models by imitating the processing of the human auditory system [4]. Despite extensive research efforts, computer algorithms based on binaural signals are not able to compete with the performance achieved by the human auditory system, and up to this point, computers are only able to perform a very restricted version of auditory scene analysis. In contrast to the human auditory system, which is remarkably robust against environmental noise and variations of acoustic conditions, computational models are usually trained for a particular acoustic scenario, and therefore, any mismatch between the training and the testing condition will decrease performance.

A powerful framework that attempts to overcome the afore- mentioned limitations by implementing concepts of ASA is the classification with missing, unreliable acoustic information [5], termed missing data (MD) classification, which is able to circumvent the mismatch between training and testing conditions. The acoustic input is first decomposed into individual time-frequency (T-F) units. Based on this two-dimensional segmentation, a so-called binary mask is used to identify whether an individual is reliable (i.e., dominated by the target source) or unreliable (i.e., dominated by noise, interfering sources or reverberation).It has been shown that such an ideal binary mask (IBM), where the assignment of reliable and unreliable T-F units is known *a priori*, can substantially improve the recognition of speech [5] and the identification of speakers [6], [7] in noisy conditions. Furthermore, it has been reported that applying an IBM to noisy speech can improve speech intelligibility in challenging acoustic scenarios for both normal hearing listeners [8], [9], as well as hearing-impaired subjects [10]. Consequently, the estimation of the ideal binary mask has been suggested to be the main goal of CASA [11].

An important aspect that is exploited by the human auditory system is the fact that it is provided with inputs from the left and the right ears. Given a complex acoustic scene, the human auditory system is able to

benefit from the spatial separation between target and interfering sources [12]. By analyzing only the signals reaching both ears, humans can detect and localize a target in the presence of up to five competing sources [13]. There are a number of computational approaches that have used binaural cues in order to estimate the ideal binary mask, either to perform robust speech recognition [14], [15], or to segregate a target source from background noise [16]. However,

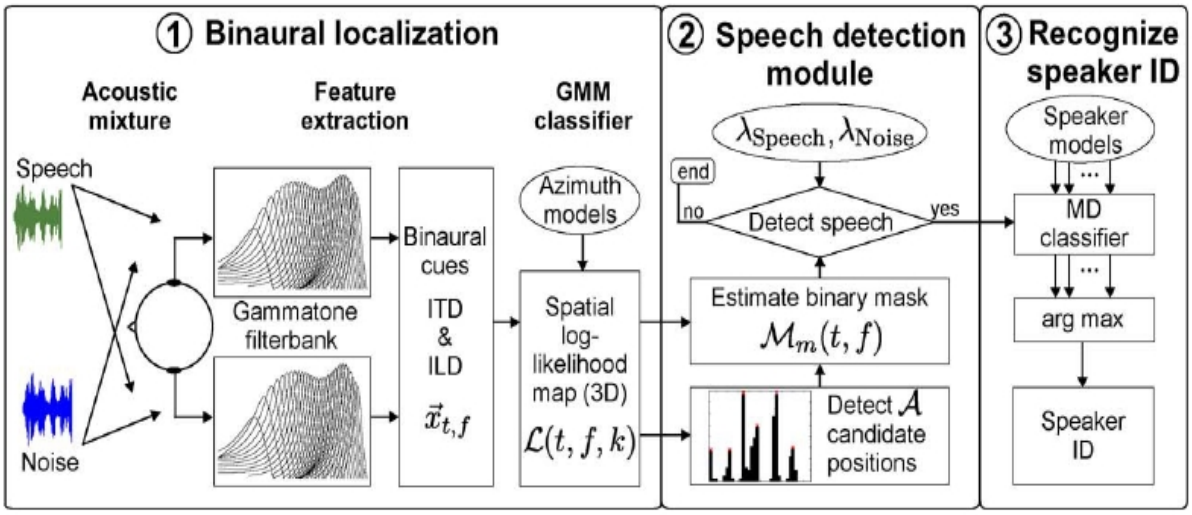


Fig. 1. Schematic diagram of the proposed binaural scene analyzer. The system is divided into three main stages: binaural localization stage, detection of speech sources, and recognition of speaker identities. See Section second for details.

An important drawback of these existing systems is that the location of the target source is assumed to be known *a priori*, which is a strong limitation for practical applications. A related area of research has focused on the localization of multiple speech sources in the presence of reverberation [17]-[19]. In these studies, all active sound sources in the acoustic scene were localized without further determining whether the source was speech or background noise, i.e., no inferences were possible about the nature of the sound sources. Thus, for a complex acoustic scene with multiple target speakers and interfering noise sources that are positioned at unknown spatial locations, it is not possible to simultaneously localize and recognize the target speakers with the aforementioned methods.

However, a wide range of applications such as hearing aids and teleconference systems require *a priori* knowledge about the azimuth location of the target sources, e.g., to steer a beam- former or to control processing parameters. Also the human auditory system is able to take advantage of *a priori* knowledge about the spatial configuration of sound sources in complex acoustic scenes. In multitalker scenarios, a significant performance gain in speech recognition has been reported when the subject's attention was directed towards the spatial location of the target talker [20]. Likewise, *a priori* knowledge about the locations of maskers in multitalker mixtures has been shown to substantially reduce the localization error of speech for humans [21]. Thus, assistive systems which are able to retrieve information about the spatial position of target speakers can potentially be used to guide the attention of human listeners.

This paper addresses the problem of jointly localizing and recognizing a known number of target speakers in adverse acoustic scenarios based on the analysis of binaural signals. For this purpose, a binaural scene analyzer is proposed that is able to simultaneously localize, detect and identify a predefined number of N speakers in the presence of reverberation and interfering noise sources that are placed at various spatial posititions. As opposed to many other cocktail party processors, a speech detection module is proposed to link the localization and the recognition stage, thus allowing the system to operate without a without *a priori* knowledge about the azimuth position of the target speakers. The proposed system builds on a previously developed binaural front-end for robust sound source localization [18], which is used to determine azimuth positions with relevant sound source activity. Based on this initial set position of target speakers can potentially be used to guide the of candidate positions, a speech detection module is presented to select azimuth positions that most likely correspond to speech sources. The final stage of the binaural scene analyzer recognizes the speaker identities of all detected speech sources. Therefore, the estimated azimuth position of the speech source is also used to select the *better ear* feature space for recognition, which aims at improving the signal-to-noise ratio (SNR) of the target speaker.

The performance of the proposed binaural scene analyzer is systematically evaluated in simulated multisource scenarios. The estimated binary mask of the proposed system is compared with two formulations of the ideal binary mask and with a binaural system proposed by Palomäki *et al.* [22]. Furthermore, speaker recognition experiments are conducted to compare speaker identification accuracy of the proposed system with the performance of MFCC-based recognizers. The remainder of the paper is organized as follows. The proposed binaural scene analyzer is described in the next section. Section 3rd contains details about baseline systems and the evaluation procedure. The experimental results are shown in Section 4th . Section 5th presents concluding remarks and summarizes the paper.

II. MODEL ARCHITECTURE

This proposed binaural scene analyzer contains of three main components, namely the binaural localization stage (1), the speech detection module (2) and the speaker recognition stage (3) .The system is shown in Fig. 1 and the individual processing stages will be described in detail in the following sections.

2.1 Binaural Localization

The binaural localization stage (1) is based on previously developed auditory front-end for robust sound source localization [18]. The acoustic input to the model is a binaural signal consisting of speech and noise sources that are randomly positioned at unknown spatial locations. The input (sampled at a rate of 16 kHz) is first split into auditory channels using a bank of Q=32 gammatone filters with center frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale[23] between 80 and 5000 Hz. More specifically, a fourth-order phase compensated gammatone filterbank [24] is used to synchronize the binaural analysis across all gammatone channels at a common time instance. Neural transduction of inner hair cells is simulated by half-wave rectification and square root compression. Afterwards, interaural time (ITD) and level differences (ILD) are independently estimated for each auditory channel using overlapping frames of 20 ms with a 10-ms shift. The ITD is estimated by detecting peaks in the normalized cross-correlation function and the ILD is derived by comparing the energy between the left and the right ear signals. These two binaural cues are combined in a two-dimensional binaural feature space:

$$\vec{x}_{t,f} = \{itd_{t,f}, ild_{t,f}\} \tag{1}$$

where *t* is the frame number and *f* indexes the gammatone channel. As shown in [18], the joint analysis of both binaural cues facilitates the disambiguation of the ITD cue by the ILD information, which is particularly beneficial in reverberant environments.

Similar to other binaural systems [16], [15], the applied localization model is based on the supervised learning of ITDs and ILDs. A noticeable difference is that the model proposed in [18] employs a multi-conditional training stage which incorporates the uncertainties of binaural cues that are caused by a variety of acoustic conditions, including changes in the source/receiver configuration, the presence of competing sound sources and the impact of reverberation. In the present study, the localization model is extended to also include different radial distances between the source and the receiver (see Section III-A for more details regarding the training).Based on the joint analysis of both binaural cues, the likelihood for each source location is determined by a Gaussian mixture model (GMM) classifier that has learned the azimuth-dependent distribution of ITDs and ILDs. Given a set of K sound source directions {$\boldsymbol{\varphi}$1,……,$\boldsymbol{\varphi}$k} that are modeled by a set of frequency-dependent GMMs$\{\lambda_{f,\varphi 1,\ldots\ldots\ldots\ldots}\lambda_{f,\varphi\kappa}\}$, a three-dimensional spatial log-likelihood map can be computed that represents the probability that the th sound source direction is active at frame and frequency channel :

$$£(t,f,k)=log\ p(\vec{x}_{t,f} \mid \lambda_{f,\varphi\kappa}) \tag{2}$$

where $p(\vec{x}_{t,f} \mid \lambda_{f,\varphi\kappa})$ is Gaussian mixture density consisting of U weighted component densities. As determined in [18], a constant GMM model complexity of U=15 Gaussian components for all gammatone channels and azimuth directions was found to give accurate localization performance. In the present study, a set of K=37 azimuths spaced by 5° within the range of [-90°,90°] is considered.

2.2 Detection of Speech Sources

The task of the speech detection module (2), as shown in Fig. 1, is to use the spatial evidence supplied by the binaural front-end to find candidate positions with relevant sound source activity. From this initial set of candidate positions, a known number of *N* sources are selected that are most likely speech by exploiting the distinct spectral characteristics of speech and noise signals. To this end, first, the evidence about a sound source location is integrated across all *Q* gammatone channels, and the most probable sound source position is used to reflect the azimuth estimate for each time frame:

$$P\hat{}_T(\mathrm{t})=\arg max_k \sum_{f=1}^{Q} £(t,f,k) \quad (3)$$

Note that this across-frequency integration of probabilities can be viewed as an implementation of the *straightness weighting* according to [25], [26], which makes the detection of sound source positions that are consistently active across multiple frequency channels more likely.To obtain a reliable estimation of active sources, all frame based azimuth estimates $P\hat{}_T$ are pooled together over the entire mixture to form an azimuth histogram *H[k]* . This implies that the sound source positions are stationary throughout the time interval over which the histogram is calculated. *H[k]* represents the number of azimuth estimates that are assigned to the *k*th sound source direction. Peaks within this histogram indicate azimuth directions with relevant sound source activity and the corresponding histogram bin indices are used to form an initial set of *A* speech source candidate positions $\mathrm{L}=(\ell_{1,\ldots\ldots\ldots},\ell_A)$. Each bin index ℓ_m corresponds to a local peak in the azimuth histogram.

Based on such a histogram, however, it is not possible to decide whether the detected activity corresponds to a speech source or to interfering noise. Nevertheless, assuming that all sources are spatially separated, the spatial information can be used to determine and isolate the contribution of individual sound sources on a T-F basis. To achieve this, the spatial log-likelihood map *£(t,f,k)* is used to estimate a binary mask $\mathcal{M}_m(t,f)$ for each of *A* the candidate positions by grouping T-F units according to common azimuth locations. More specifically, for each T-F unit the most likely position among all ***m=1,…….,A*** candidate positions is determined, and the individual T-F unit is added to the corresponding mask:

$$\mathcal{M}_m(t,f)=\begin{cases}1 & if\ m=argmax_{k\epsilon\mathcal{L}}\,\mathcal{L}(t,f,k)\\ 0 & otherwise\end{cases} \quad (4)$$

Rather than considering all K possible sound source directions, the candidate selection effectively reduces the number of alternatives per T-F unit, which results in a more dense binary mask. Based on this mask, missing data classification [5] is performed to decide whether the corresponding source type is speech or noise. Our prior work has shown that the mean absolute deviation of the smoothed envelope is a good descriptor for detecting speech sources in the presence of noise and reverberation [27]. In order to compute this feature, first a smoothed envelope e_f is obtained by low-pass filtering the half-wave rectified output of the *f*th gammatone channel with a time constant of 10 ms. Then, the mean absolute deviation of the smoothed envelope e_f is computed over B adjacent samples with a shift of R samples:

$$F(t,f)\frac{1}{B}\sum_{b=0}^{B-1} \mid e_f(tR+b)-\bar{}e_f \mid \quad (5)$$

Where $\bar{}e_f$ refers to the mean of the envelope of the *t* th frame. The feature reflects the amount of fluctuation and its magnitude lower for speech-dominant T-F units compared to units that corrupted by noise. Thus, it is possible to apply *bounded marginalization* where the true value of the unreliable feature components is bounded between zero and the observed feature magnitude [5], [28], [29]. Signals of the left and the right ears are averaged prior to feature extraction. Similar to the binaural front-end, the processing is based on 20-ms frames with a shift of 10 ms. For classification two GMMs with 32 Gaussian components and diagonal covariance matrices are trained.

The probable distribution of the feature space *F(t,f)* that is extracted separately for speech and noise files. The first GMM, denoted as speech model $\lambda_{Speec\ h}$, is trained with features based on a large pool of monaural speech files selected from the speech separation challenge (SSC) database [30]. The second GMM, termed noise model λ_{Noise}, reflects the feature distribution of all types of noise files drawn from the NOISEX database[31].The GMMs are initialized by 20 iterations of the k means clustering algorithm [32] and afterwards refined by the EM algorithm [33] using a stopping criterion $1e^{-5}$ of with a maximum of 300 iterations. About 29 minutes of training material is used for each GMM.

To compensate for the mismatch between training ($\lambda_{Speec\ h}$, and λ_{Noise}, are trained with clean signals) and testing (the speech detection module is applied in noisy and reverberant conditions), a missing data compatible normalization scheme is employed. Therefore, a frequency-dependent compensation factor is derived by computing the mean of the most intense feature values that are classified as being reliable by the estimated binary mask [22].

Given the binary mask $\mathcal{M}_m(t,f)$, the feature space , *F(t,f)* and the trained speech and noise models ($\lambda_{Speec\ h}$, and λ_{Noise} ,), the log-likelihood ratio $\mathcal{P}_m$ reflecting the evidence for the *m* th speech source candidate can be determined by:

$$\mathcal{P}_m = log\left(\frac{p(\mathcal{F}|\lambda_{Speec\ h})}{p(\mathcal{F}|\lambda_{Noise})}\right) \quad (6)$$

A speech source is detected if the log-likelihood ratio is larger than a predefined threshold θ:

$$\mathcal{P}_m \begin{cases} \geq \theta, & accept\ \lambda_{Speec\ h} \\ < \theta, & reject\ \lambda_{Speec\ h} \end{cases} \tag{7}$$

Selecting the optimal decision threshold is a nontrivial task because it is influenced by a variety of parameters; among them the number of speech and noise sources in the acoustic mixture, the SNR between all sources and the amount of reverberation. In preliminary experiments we found that a decision threshold of $\theta = 0$ performed well for a wide range of acoustic scenarios. Based on this criterion, all active sound sources are classified to either speech or noise. After classification a set of $\widehat{\mathrm{N}}$ loglikelihood ratios$\{p1^{Speec\ h}, \ldots\ldots\ldots\ldots\ldots p\widehat{\mathrm{N}}^{Speec\ h}\}$ is available that specifies the evidence of all detected speech sources. Moreover, a new set of histogram bin indices $Ł^{Speec\ h}=\{\ell 1^{Speec\ h}, \ldots\ldots\ldots\ldots\ldots \ell\widehat{\mathrm{N}}^{Speec\ h}\}$ is available, which is a subset of Ł , and reflects the individual bin positions of all detected speech sources in the azimuth histogram.

Reflections and the interaction of multiple competing sound sources can cause the azimuth histogram to have numerous local is $\widehat{N}$ might be larger than the number of *a priori* known speech sources N . Thus, the final step is to select the *N* most likely speech sources. Instead of using the evidence from the missing data classifier directly for selection, we found that it is advantageous to apply an azimuth dependent weight to the log-likelihood ratio of each detected speech source to account for the fact that speech sources that are more frequently represented in the azimuth histogram are more likely to reflect the real position of the speech sources [27]. The applied weight reflects the *a priori* probability that the corresponding source was active in the acoustic scene, and is approximated by the normalized histogram which is azimuth histogram. The weighted log-likelihood ratio for the n^{th} speech source is given by:

$$\mathcal{P}_n^{Speec\ h,W} = \mathcal{P}_n^{Speec\ h} + log\left(\frac{H\left[\ell_n^{Speec\ h}\right]}{\Sigma_k H[k]}\right) \tag{8}$$

Azimuth weight

Finally, the set of weighted log-likelihood ratios of all detected speech sources is rearranged in descending order:

$$\{p1^{Speec\ h,w} \geq p2^{Speec\ h,w} \geq, \ldots\ldots\ldots\ldots\ldots\ldots\ldots \geq p_{\widehat{\mathrm{N}}}{}^{Speec\ h,w}\} \tag{9}$$

The azimuth locations corresponding to the highest values are selected to represent the estimated speech source positions. When using the frame-based azimuth estimates according to (3) for the initial selection of source candidate positions, it is required that a sufficiently large number of frames is dominated by the target the target sources which should be detected. However, in conditions where the SNR between the target speakers and the interfering noise sources is very low, or even negative, very few target source dominated frames may be found. Thus, when $\widehat{N} < N$, the histogram of the frame-based azimuth estimates $\widehat{\mathrm{P}_{\mathrm{T}}}$ was apparently dominated by locations corresponding to noise sources, and the histogram did not reflect the locations of all present speech sources. Indeed, it has been shown that spectro-temporal regions dominated by speech tend to be sparse in the presence of noise [34]. Thus, whenever $\widehat{N} < N$, the azimuth histogram *H[k]* is recomputed using the azimuth estimates on a T-F basis:

$$\widehat{\mathrm{P}_{\mathrm{T,F}}}(\mathrm{t},\mathrm{f}) = \arg\max_k \mathcal{L}(t,f,k) \tag{10}$$

Again, all local peaks within this histogram are considered as initial speech source candidates, and the missing data masks corresponding to these locations are estimated and fed to the missing data classifier to determine the most likely speech source positions [involving the aforementioned steps (4)-(9)].The rationale behind (10) is that speech source positions that be were not resolved on a frame-by-frame basis can potentially be recovered when using azimuth estimates on a T-F level. In Section IV-A, the impact of alternative methods to select the set of speech source candidate positions is analyzed in order to justify the proposed frame-based selection of speech source candidates with the possibility to switch to time-frequency-based processing.

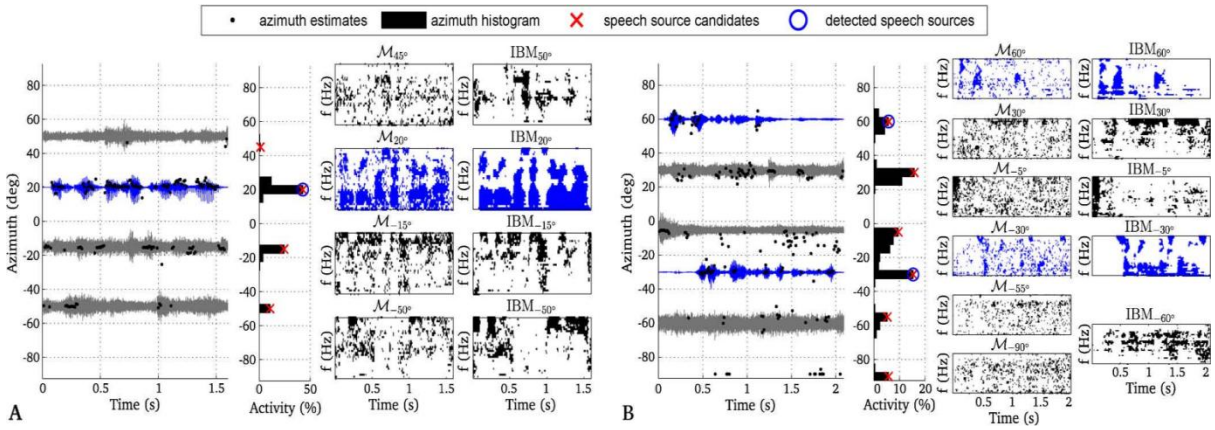


Fig. 2. Demonstration of the speech detection module for two different acoustic scenes.

The proposed speech detection module is illustrated in Fig. 2 for two different acoustic scenes. Fig. 2(A) shows the detection of one speech source in the presence of three factory noise sources in an anechoic room. Despite the presence of four competing sources, the estimated binary masks are quite similar to the ideal binary masks based on the *a priori* SNR. Fig. 2(B) presents the detection of two speech sources in the presence of three factory noise sources in a reverberant room(T_{60} = 0.29s).In comparison to the anechoic scenario with one target source, the estimated binary masks of the two target sources are more noisy due to the impact of reverberation.

2.3 Automatic Speaker Recognition

The final stage (3) of the proposed binaural scene analyzer (see Fig. 1) has the function of recognizing the speaker identity of the detected speech sources from a set of stored speaker models. For this purpose, a second missing data classifier based on bounded marginalization is supplied with the binary masks that were estimated by the speech detection module (see Section II-B). The recognition of speakers is performed with spectral features reflecting the energy of individual frequency channels [5]. Therefore, a map of auditory nerve firing rates, a so called *ratemap*, is computed by averaging the smoothed envelope e_f(see Section II-B) over B adjacent samples with a shift of R samples and subsequent cube-root compression

$$\mathrm{R(t,f)} = \left(\frac{1}{B}\sum_{b=0}^{B-1} e_f(tR+b)\right)^{\frac{1}{3}} \tag{11}$$

Speaker models are represented by 128-mixture GMMs with diagonal covariance matrices. In comparison to a conventional GMM-based missing data recognizer, we recently found that the combination of missing data recognition with universal background model (UBM)-based adaptation of speaker models [35] yields substantial improvements in highly non stationary noise scenarios [7]. However, in order to apply this scheme to reverberant multi-source environments, a modification is required to account for the mismatch between the speaker models trained with monaural and anechoic speech, and the observed spectral features that are affected by HRTF filtering and reverberation. Similar to the speech detection module, a missing data compatible normalization scheme is required. The normalization scheme proposed in [22] was developed in the context of automatic speech recognition and applies a spectral normalization factor that is independently derived for each frequency channel. In contrast to automatic speech recognition, which is generally speaker-independent, an automatic speaker identification system exploits the frequency-dependent spectral variations across different speakers. We found that it is beneficial in terms of speaker recognition performance to average the normalization factor proposed in [22] across adjacent frequency channels prior to normalization. In this way, a similar normalization factor is applied to neighboring frequency channels and differences between adjacent channels will be related to speaker-specific variations. A sliding triangular window of size 7 is used for averaging the normalization factor (these parameters were derived empirically based on pilot experiments).

For the adaptation of speaker models, two gender-dependent UBMs are used to represent the speaker-independent distribution of the ratemap feature. The two UBM models are initialized with 20 iterations of the -means clustering algorithm [32] and further trained with the EM algorithm [33] using a stopping criterion $1e^{-5}$ of at a maximum of 300 iterations. Speaker-dependent models are obtained by adapting the well-trained UBM parameters to the speaker-dependent speech material. Therefore, first the gender selection is performed by selecting the UBM which shows the highest probabilistic alignment with the speaker-dependent material. Second, as suggested in [35], only the mean vectors of the UBM are adapted using a relevance factor of 16.

In order to benefit from the fact that the binaural scene analyzer is provided with two acoustic signals from both ears, the estimated positions of speech sources are utilized to implement a better-ear selection of the feature space. Therefore, ratemaps are always computed for both the left R_1(t,f) and the right ear R_r(t,f) signals . For recognition, the ratemap based on the ear signal that is closest to the estimated azimuth position of the

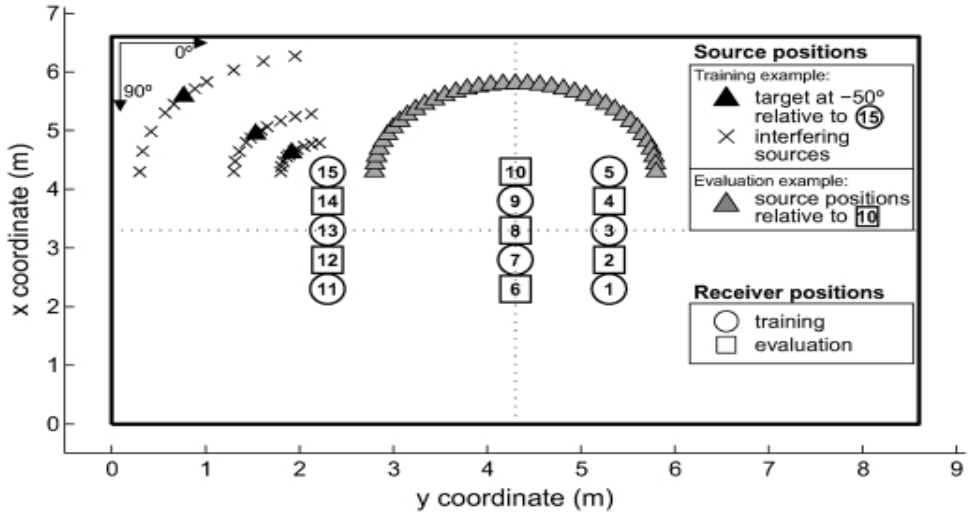


Fig. 3. Schematic diagram of the room dimensions with all receiver positions used for training (circles) and evaluation (squares). Note that the training stage of the localization model incorporated three radial distances (0.5 m, 1 m, and 2 m) between the receiver and the target positions as exemplarily shown for receiver position 15, which were different from the radial distance (1.5 m) used for evaluation. See Section III-A for details.

corresponding speech source is selected individually for each detected speech source:

$$\mathcal{R}_m(t,f) = \begin{cases} \mathcal{R}_1(t,f), \\ \mathcal{R}_r(t,f), \\ \frac{1}{2}\left(\left(\mathcal{R}_1(t,f) + \mathcal{R}_r(t,f)\right)\right) \end{cases} \tag{12}$$

This approach aims at increasing the SNR between the target and interfering sources, and the underlying effect is referred to as the *better ear* effect [36].

III. EVALUATION SET UP

3.1 Acoustic Mixtures

Acoustic sources were simulated by convolving monaural audio files with binaural room impulse responses (BRIRs). BRIRs were constructed by combining head related transfer functions (HRTFs) of a KEMAR artificial head taken from the MIT database [37] with room impulse responses (RIRs) that were simulated according to the image-source model [38]. More specifically, the *roomsim* simulation software [39] was used for that purpose. The receiver (KEMAR) was placed at various positions in a simulated room of dimensions $6.6 * 8.6 * 3$ m at 1.75 above the ground, as shown in Fig. 3.

The binaural localization model was trained with BRIRs corresponding to eight training positions (different from those used for evaluation). To incorporate the same amount of uncertainty to the binaural cues of all gammatone channels, we intentionally chose a frequency-independent reverberation time of $T_{60} = 0.5$ s for all training positions. It has been shown that this training enables the localization model to generalize to unseen absorption characteristics [18]. Note that a different, frequency-dependent absorption characteristic is used for computation. The training of the binaural model also incorporated the effect of interfering sources. This is shown in Fig. 3. for the training position (15). In order to train the binaural model for one particular sound source direction, the training consists target source placed at the corresponding azimuth (denoted by ▲) and an interfering source (indicated by X) positioned at $\pm 5^\circ$, $\pm 10^\circ$, $\pm 20^\circ$, $\pm 30^\circ$ and $\pm 40^\circ$ with respect to the target azimuth. In addition to the training procedure explained in [18], the multi-conditional training is extended to consider three different radial distances (0.5 m, 1 m, and 2 m) between the source and the receiver.

For evaluation the receiver was randomly placed at seven evaluation positions using a radial distance of 1.5 m. To systematically evaluate the impact of reverberation, the surface *Acoustic plaster* was selected for all room boundaries within the room simulation software [39] to create a specific, frequency-dependent

absorption characteristic. We can note that this absorption characteristic was different from the one used to train the localization model in system. Speech and noise sources (diffferent from the material used to train the speech detection module) were randomly positioned within the azimuth range of 90 [-90˚, 90˚] while having an angular distance of at least 15 to the nearest source. A set of 1400, four-source mixtures (one speech source) and 600, five-source mixtures (two speech sources) were produced for each SNR condition. Mixtures had an average length of 1.83 s. The SNR was adjusted by comparing the A-weighted energy of all binaural speech sources with the A-weighted energy of all binaural noise sources. This weighting was applied to ensure that the SNR is adjusted in the frequency range that is relevant for speech. The design of the A-weighting filter was implemented according to [40]. To prevent that the energy of speech is underestimated due to silent parts, an energy-based voice activity detector (VAD) was used. A frame was considered to contain relevant speech activity, if its energy level was within 40 dB of the global maximum. The level between multiple speech or noise sources was always set equal. For a given multi-source mixture, the localization error in degrees was evaluated by comparing the positions of the detected speech sources to their real positions.

Speaker recognition performance was evaluated on a closed set of speakers that were randomly selected from the SSC database [30]. The SSC database consists of 17 000 clean utterances spoken by 34 speakers (18 males and 16 females). To ensure that there is no overlap between the speech material used for training and testing, the SSC database was randomly split into two equal sized sets consisting of 8500 files (250 sentences per speakers). These two are very advantageous according to this system. The first half was used to train the two gender dependent UBMs. Also, 950 sentences (about 29 minutes) were randomly selected from the first half to train the speech model which is denoted by $\lambda_{Speec\ h}$ of the speech detection module. The second half of the SSC database was used to perform the speaker recognition experiments reported in Section IV, involving the training of speaker- specific models using UBM adaptation and the evaluation of the speaker identification accuracy because the amount of available speech material is often a limitation for practical applications, the speech material was restricted to 25 randomly selected sentences per speaker. For each speaker, 18 sentences were randomly chosen to train the speaker model and the remaining 7 sentences were used for evaluation. Because this randomized selection of training and testing material will to some extent influence the evaluated speaker identification accuracy, results are reported as the mean identification accuracy over a series of 20 simulations, each containing a new set of randomly selected speakers. Note that the speaker identification accuracy was measured on an utterance level.

3.2 Baseline Systems

To serve as a baseline for recognition performance, a conventional robust speaker recognition system was trained with a feature vector contains 13 static MFCC coefficients including the zeroth-order coefficient and first-order temporal derivatives (a total of 26 features). The static MFCC coefficients were computed using the RASTAMAT toolbox [41]. Parameters1 were chosen to regenerate MFCC coefficients according to the hidden markov models toolkit (HTK). The delta coefficients were computed using a first-order orthogonal polynomial fit over a window of five frames [42]. For improved robustness, cepstral mean and variance normalization (CMVN) was performed, where the feature statistics are measured over the duration of one utterance [43], [44].

The first method, named *MFCC Mono*, extracted the MFCC coefficients by averaging the signals of the left and the right ear. In addition, a recognizer was implemented that calculated the MFCC feature vector, as explained above, for both the left and the right ear signals. Based on the binaural analysis, the feature vector with the higher SNR (the *better ear*) was selected for recognition according to the computed location of each target speech source, individually. This algorithm is referred to as *MFCC Binaural*. The third MFCC-based recognizer, denoted by *MFCC Binaural NR*, combined the previously described better-ear selection and a noise reduction stage. In a previous study [7], we analyzed the impact of a variety of noise computation and noise reduction schemes on speaker identification performance. Based on these findings, noise reduction is performed prior to MFCC extraction by recursively averaging the noise floor [45] in combination with the *MMSE log-STSA* gain function [46]. The noise floor computation and the noise reduction is applied after the speech detection module and is performed independently for the left and the right ear signals.

Similar to the MD-based recognizer, speaker models of all MFCC-based recognizers are represented by 128-mixture GMMs and were adapted from two gender-dependent UBMs. Recognition of two target speakers is performed by accumulating the frame-based likelihoods over the entire test sequence and selecting the two most likely speaker identities.

Finally, a comparison is made with a recently proposed co-channel speaker identification system based on adapted GMMs [47]. This approach, denoted as *MFCC Co-channel*, combines frame-level likelihood scores and a Kullback-Leibler divergence (KLD) distance measure to find the most likely speaker identity on a frame-by-frame basis in co-channel scenarios. A UBM with 128 Gaussian components is trained with MFCC coefficients extracted from two-talker mixtures. For training, two-talker mixtures are created from the first half of the SSC database by mixing two sentences from different speakers at one of these following seven signal-to-

signal ratio (SSR) levels which are shown above.

A set of 8500 co-channel mixtures is created, giving a total of 59 500 audio files for UBM training. Speaker-dependent GMMs are adapted from the UBM by mixing 18 randomly selected training sentences for each speaker (see Section III-A for details) with 18 files from other speakers, again at different SSR levels. Because multiple talkers are always set to have equal power in the experiments, only the following SSR levels are considered for adaptation: $\{0,\infty\}$ dB.

3.3 Ideal Binary Mask

To compute upper performance limit of missing data recognition systems, an ideal binary mask is commonly introduced that represents the ideal segmentation of the spectral feature space according to the contribution of all occurring sound sources. Research work related to recognition tasks in noisy conditions often utilize the ideal binary mask based on the *a priori* SNR between the target and the noise source [5]. Note that the SNR-based ideal binary mask, denoted as *IBM SNR*, is only able to segregate the target signal from the background noise, but the effect of reverberation is not taken into account. Another formulation of the ideal binary mask is to select only those T-F units where the spectral energy of the noisy and reverberated speech is within 3 dB of the spectral energy of the clean target source [5], [14], taking into account both the effect of background noise and the impact of reverberation. The ideal binary mask based on this spectral criterion will be referred to as *IBM SPEC*. To create this mask for a given binaural mixture, the observed spectral energy of the noisy and reverberated speech signal is compared with the spectral energy of the clean speech signal, that has been convolved with the same HRTF but in anechoic conditions. In this way, the azimuth-dependent HRTF filtering does not bias the mask computation. Both definitions of the ideal binary mask will be used to evaluate the mask estimation performance of the proposed binaural front-end in the experimental section.

IV. EXPERIMENTS

We performed a series of experiments about the localization and speaker recognition to evaluate the proposed binaural scene analyzer in simulated adverse acoustic conditions. The first two experiments are aimed at evaluating the computed localization information of the proposed system. Whereas the first experiment computes the ability of the speech detection module to localize the azimuth of speakers in multi-source scenarios, the second experiment analyzes the mask estimation performance of the binaural front-end. Therefore, the computed binary mask is compared with two different formulations of the ideal binary mask and with the model proposed by Palomäki *et al.* [14]. Both the computed location of target speakers and the computed binary masks are used in the third and fourth experiment to identify the identity of speakers in complex acoustic scenes. More specifically, the third experiment is using a reduced set of ten speakers to study the influence of the number of interfering noise sources on speaker recognition performance. Furthermore, the performance of the proposed system is compared with several baseline systems based on MFCC coefficients (see Section III-B). Based on this comparison, the best performing baseline systems and the proposed method are used active target speakers using the full set of 34 speakers. In the last experiment, the final aim of this study is addressed by jointly analyzing the combined localization and speaker recognition performance of the proposed method using a confusion matrix.

TABLE I

SNR-Dependent localization error of speech sources in degrees for binaural mixtures consisting of one and two speakers in the presence of interfering noise sources and reverberation

$T_{60} = 0.29$ s	Candidate selection	SNR in dBA (factory noise) −5	0	5	10	20	Mean
1 speaker, 1 noise source	$\hat{P}_{\mathrm{TF}}$	23.7	10.9	3.6	1.7	0.5	8.1
	$\hat{P}_{\mathrm{T}}$	17.0	2.2	0.6	0.1	0.0	4.0
	Proposed	12.0	1.6	0.4	0.1	0.0	2.8
1 speaker, 2 noise sources	$\hat{P}_{\mathrm{TF}}$	19.5	6.1	1.7	0.4	0.6	5.7
	$\hat{P}_{\mathrm{T}}$	12.2	1.4	0.4	0.1	0.0	2.8
	Proposed	11.4	1.3	0.4	0.1	0.0	2.6
1 speaker, 3 noise sources	$\hat{P}_{\mathrm{TF}}$	16.7	4.1	1.1	0.5	0.4	4.6
	$\hat{P}_{\mathrm{T}}$	9.8	1.5	0.3	0.1	0.1	2.4
	Proposed	9.2	1.5	0.3	0.1	0.1	2.2
2 speakers, 1 noise source	$\hat{P}_{\mathrm{TF}}$	31.0	24.2	15.5	9.4	3.4	16.7
	$\hat{P}_{\mathrm{T}}$	32.3	12.0	3.1	1.5	0.9	10.0
	Proposed	27.8	12.1	3.0	1.5	0.9	9.1
2 speakers, 2 noise sources	$\hat{P}_{\mathrm{TF}}$	31.5	21.2	13.5	7.3	3.6	15.4
	$\hat{P}_{\mathrm{T}}$	29.4	12.7	4.6	1.5	1.0	9.8
	Proposed	27.4	12.6	4.6	1.5	1.0	9.4
2 speakers, 3 noise sources	$\hat{P}_{\mathrm{TF}}$	29.2	18.1	11.7	6.9	4.1	14.0
	$\hat{P}_{\mathrm{T}}$	26.2	11.5	4.4	2.3	0.4	9.0
	Proposed	25.8	11.4	4.4	2.3	0.4	8.9

4.1 Experiment 1: Speaker Localization Performance

This experiment is analyzing the ability of the speech detection module to evaluating the azimuth position of a known number of speech sources from a set of candidate positions. Furthermore, the influence of the speech source candidate selection on speech detection performance is systematically investigated. Therefore, we compare the proposed candidate selection as described in Section II-B with two alternative methods.

The SNR-dependent localization error in degrees of the detected speech sources is shown in Table I for all three candidate selection methods. It can be seen that the T-F-based selection of speech source candidates generates the highest error rates. When using the frame-based selection , a significant improvement is obtained. This improvement can be attributed to the fact that the frame-based candidate selection integrates evidence of sound source activity across all frequency channels, which effectively increases the reliability of the resulting localization estimate. As a result, the number of candidate positions is reduced, which consequently reduces the number of alternatives per T-F unit in (4), thus increasing the density of the computed binary mask which allows for a more accurate detection of the speech sources.

The proposed candidate selection, which combines both the frame-based and the T-F-based selection, can further improve the localization performance at low SNRs. In particular, in conditions with one interfering noise source, it appears that switching from frame-based to T-F-based processing can to some extent recover the position of speech sources, therefore improving the localization performance by about 5°. Regarding mixtures with one speaker, the average localization error is below 3° for SNRs as low as 0°dBA. When two speakers are concurrently talking, the azimuth of both speakers is computed within 5°accuracy for SNRs as low as 5°dBA. At lower SNRs, the error is remarkable increased.

The general trend that the localization error decreases with increasing number of noise sources can be attributed to the SNR definition explained in Section III-A, which compares the overall energy of all speech sources with the energy of all noise sources. With increasing number of noise sources, the noise energy is distributed across multiple directions, which increases the relative localization dominance of the speech source, thus allowing for a more accurate prediction of its azimuth position.

4.2 Experiment 2: Evaluation of the IBM Estimated by the Binaural Front-End

This second experiment is used to check the ability of the binaural front-end to compute the ideal binary mask. The quality is systematically evaluated in respect of receiver operating characteristics (ROC) analysis [48]. For this analysis, the computed ideal binary mask is compared with the ideal binary mask by calculating the percentage of correctly identified T-F units which are dominated by the target signal (true positive rate) and the percentage of misclassified T-F units which are dominated by interfering sources (false positive rate). For comparing this, we also provide the difference between the true positive rate and the false positive rate, because it has been shown that this metric is highly correlated with human speech intelligibility [49]. For the ROC analysis, we selected the *IBM SPEC* to represent the reference mask, which accounts for both the effect of interfering noise and reverberation. Moreover, the percentage of labeled T-F units is reported to shows the amount of information that is available to the MD classifier. In addition to the ROC analysis, the corresponding speaker identification accuracy for a set of ten speakers is provided to evaluate the implication of the true positive rate and the false positive rate on speaker recognition performance.

We compared the performance of the proposed mask computation technique with two definitions of the ideal binary mask (see Section III-C) and with a binaural front-end proposed by Palomäki *et al.* [14]. In addition, to study the influence of the speech detection module on mask estimation performance, the proposed method is supplied with *a priori* knowledge about the azimuth positions of the target and the interfering sources.The model proposed in [14] extracts both ITD and ILD cues in individual frequency channels. The ITD cue is warped by a table lookup to its corresponding azimuth and subsequently used to group T-F units according to common azimuth. The required azimuth locations of the target and interfering sources are provided by the speech detection module (see Section II-B).

TABLE II
Mask estimation performance (True positive (TP) rate, False positive (FP) rate, TP-FP rate and the number of labeled T-F Units) & Speaker identification (SID) accuracy in % for various methods.

T_{60}	Methods	%	SNR in dBA −5	0	5	Mean
0 s	IBM SPEC	TP rate	100	100	100	100
		FP rate	0	0	0	0
		TP-FP rate	100	100	100	100
		T-F units	26.1	37.2	48.8	37.4
		SID accuracy	97.1	98	98.4	97.8
	IBM SNR	TP rate	64.9	72.3	78.2	71.8
		FP rate	0	0	0	0
		TP-FP rate	64.9	72.3	78.2	71.8
		T-F units	16.9	26.9	38.2	27.3
		SID accuracy	92.5	97.8	98.9	96.4
	Proposed *a priori*	TP rate	59	66.3	72.5	65.9
		FP rate	2.4	3.1	3.9	3.1
		TP-FP rate	56.6	63.2	68.6	62.8
		T-F units	17.2	26.6	37.4	27.1
		SID accuracy	90.5	97.8	98.9	95.7
	Proposed	TP rate	54.8	61.6	68.1	61.5
		FP rate	3.2	3.7	4.1	3.7
		TP-FP rate	51.6	57.9	64.0	57.8
		T-F units	16.6	25.2	35.4	25.8
		SID accuracy	84.8	94.9	97.6	92.4
	Palomäki ITD	TP rate	45	53.8	61	53.3
		FP rate	4.5	6.4	8.3	6.4
		TP-FP rate	40.5	47.4	52.7	46.9
		T-F units	15.1	24	34.1	24.4
		SID accuracy	79.5	93.4	97.6	90.2
	Palomäki ITD & ILD	TP rate	42.5	49.2	54.5	48.7
		FP rate	3.5	4.9	6.3	4.9
		TP-FP rate	39.0	44.3	48.2	43.8
		T-F units	13.7	21.4	29.8	21.6
		SID accuracy	67.1	80.9	90.3	79.4
0.29 s	IBM SPEC	TP rate	100	100	100	100
		FP rate	0	0	0	0
		TP-FP rate	100	100	100	100
		T-F units	18.4	27.8	38.1	28.1
		SID accuracy	95.3	97.5	97.7	96.8
	IBM SNR	TP rate	68.6	77.1	83.5	76.4
		FP rate	5.9	10.1	16	10.7
		TP-FP rate	62.7	67.0	67.5	65.7
		T-F units	17.5	28.7	41.7	29.3
		SID accuracy	85.4	94.8	97.3	92.5
	Proposed *a priori*	TP rate	58.7	64.5	69.9	64.3
		FP rate	12.6	14.7	17.8	15
		TP-FP rate	46.1	49.8	52.1	49.3
		T-F units	21.1	28.6	37.6	29.1
		SID accuracy	78.9	90.9	95.6	88.5
	Proposed	TP rate	52.5	55	58.4	55.3
		FP rate	12.2	12.2	13.5	12.7
		TP-FP rate	40.3	42.8	44.9	42.6
		T-F units	19.7	24.1	30.6	24.8
		SID accuracy	74.5	88.1	93.9	85.5
	Palomäki ITD	TP rate	37.5	41.9	45.7	41.7
		FP rate	12.2	13	14.7	13.3
		TP-FP rate	25.3	28.9	31.0	28.4
		T-F units	16.8	21	26.5	21.5
		SID accuracy	61.9	80.6	89.3	77.3
	Palomäki ITD & ILD	TP rate	34.7	37	38.8	36.8
		FP rate	11.2	11.6	13	11.9
		TP-FP rate	23.5	25.4	25.8	24.9
		T-F units	15.6	18.7	22.8	19
		SID accuracy	47.1	61.8	75.3	61.4

ILD cue is used to remove T-F units from the estimated binary mask where the ILD estimate is not consistent with the azimuth of the sound source, which is derived from the ITD analysis. The expected azimuth-specific ILD template is precomputed for all frequency channels above 2800 Hz. We computed two variants of the model, first the combined ITD and ILD analysis denoted as *Palomäki ITD & ILD*, and *Palomäki ITD* which solely relies on ITD analysis. Note that the original model proposed in [14] consist an inhibition mechanism that

emphasizes acoustic onsets. Whereas this inhibition might be beneficial for the azimuth estimation of sound sources on an utterance level (as indicated in Fig. 5 in [14]), preliminary tests shown, however, that the resulting mask was very sparse and the model performed best in terms of speaker identification performance when the inhibition mechanism was switched off by setting the inhibition gain to zero. Apart from this modification, all other model parameters were chosen according to the recommendations of the authors [14].

The computation of all tested mask estimation methods is shown in Table II for binaural mixtures with one target speaker and one interfering factory noise source. The upper and the lower half of the table present results achieved in anechoic($T_{60} = 0\ s$) and reverberant ($T_{60} = 0.29\ s$) conditions. When comparing the model proposed by *Palomäki et al*. with and without ILD constraint, we can see that the model with ILD constraint produces a lower false positive rate (FP rate) in both anechoic and reverberant conditions, but at the same time, the true positive rate (TP rate) is noticeably lower, which consequently limits speaker identification performance. Especially in the reverberant condition, speaker identification accuracy of *Palomäki ITD & ILD* is on average 15.9% below *Palomäki ITD*. We believe that these results can be explained by the employed ITD look-up table and the precalculated ILD template, which are both trained with a single source in anechoic conditions. Such a training imposes very strict constraints on the expected ITDs and ILDs. However, it has been shown that binaural cues that are associated with a target source based on the presence of interfering sources and their relative strength to the target [16]. Reverberation has a severe effect on the ILD cue [50], [51] also, which will cause the ILD constraint to remove many T-F units from the mask, although the underlying template function does not match with the acoustic condition in which the model is applied.

This proposed method achieves significantly higher TP rates and lower FP rates in comparison with *Palomäki ITD*. Whereas speaker recognition performance of both methods is comparable in the anechoic condition, this proposed method substantially outperforms *Palomäki ITD* in the presence of reverberation. Especially at lower SNRs, the speaker identification performance of the proposed model is about 12% above the one by *Palomäki ITD*. In contrast to the model of Palomäki *et al.*, The proposed binaural front-end is designed to operate in a variety of acoustic conditions, including reverberation and multi-source scenarios, due to the multi-conditional training.

When we replace the speech detection module in the proposed model with *a priori* knowledge about the locations of the target and the interfering source, it can be seen that performance is quite similar in terms of TP rates and FP rates for SNRs as low as 0 as dBA. This suggests that the speech detection module is able to robustly determine the location of the target source. Only for negative SNRs, more substantial differences can be observed, especially in terms of speaker identification accuracy. This gap can be described by the increased error rate of the speech detection module at negative SNRs, as reported in Table I.At last, we compare the two formulations of the ideal binary mask. It is very interesting to note that while the mask produced by *IBM SPEC* generally consists of more T-F units than *IBM SNR* in the anechoic condition, the mask is more sparse in reverberant conditions.Nevertheless, *IBM SPEC* consistently outperforms the IBM based on the *a priori* SNR in all experimental conditions. This denotes that in order to improve on existing mask estimation techniques, the effect of reverberation should be taken into account to reduce the degrading effect of spectral variations that are due to strong reflections.

[2]Both the mapping function and the ILD template were derived for the same HRTFs used in our experiments. Note that training these functions with reverberant HRTFs did not improve the performance of the model.

4.3 Experiment 3: Speaker Identification Depending on the Number of Interfering Noise Sources

This experiment compares the speaker identification performance of the binaural scene analyzer with a number of MFCC-based recognizers using a reduced set of ten speakers. Furthermore, the influence of the number of interfering noise sources is investigated.

The average speaker identification accuracy for one target speaker in reverberant conditions is ($T_{60} = 0.29\ s$) in Fig. 4. Panels (A)-(C) show performance based on the number of interfering sources, ranging from (A) one to (C) three concurrently active factory noise sources that are randomly placed at different spatial locations. According to the expectation, the speaker identification accuracy decreases with decreasing SNR for all methods. The performance of the MFCC-based recognizer *MFCC Mono* rapidly deteriorates with decreasing SNR. The system *MFCC Binaural*, which selects the better ear feature space according to the computed location of the target speaker, provides a substantial advantage over the monaural MFCC recognizer. This improvement can be in the range of 20% at lower SNRs. A possible discussion may be that the better ear signal has a better SNR than the monaural signal, in addition, there is less spectral distortion caused by the head shadow. We found that the advantage of *MFCC Binaural* based on the spatial separation between the target and the interfering noise sources and increases with increasing spatial separation. An additional performance gain is obtained by *MFCC Binaural NR*, where noise reduction is applied prior to MFCC extraction. This improvement is quite small for scenarios with one interfering noise source and moderately increases when two or three noise

sources are present concurrently.

Again, the proposed system *MD Proposed* is outperforming the best MFCC-based recognizer in respect of speaker identification accuracy, especially at low SNRs. In respect of acoustic mixtures with one interfering noise source, the performance of the proposed system is close to the system *MD IBM SNR* that used *a priori* SNR information, which denotes that the computed binary mask that is provided by the binaural front-end is of high accuracy. The proposed system shows a stronger dependency on the number of interfering noise sources than MFCC-based recognizers. In general, performance decreases with increasing number of noise

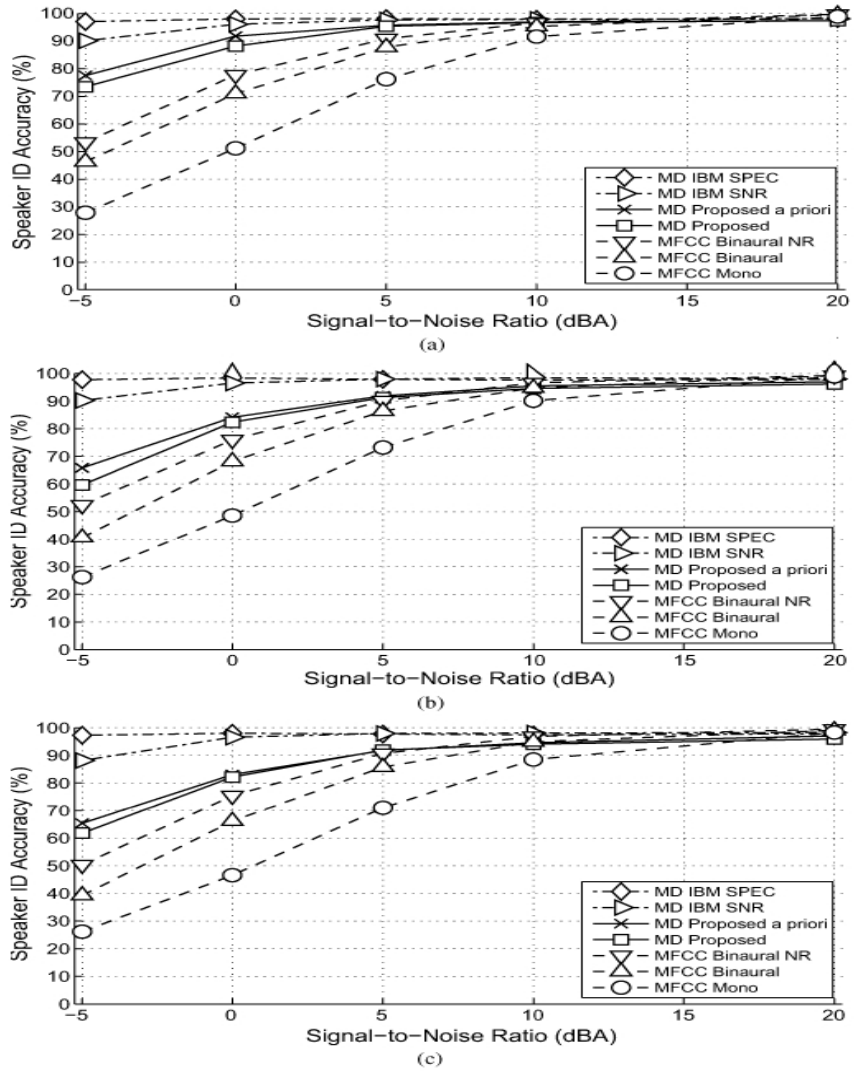
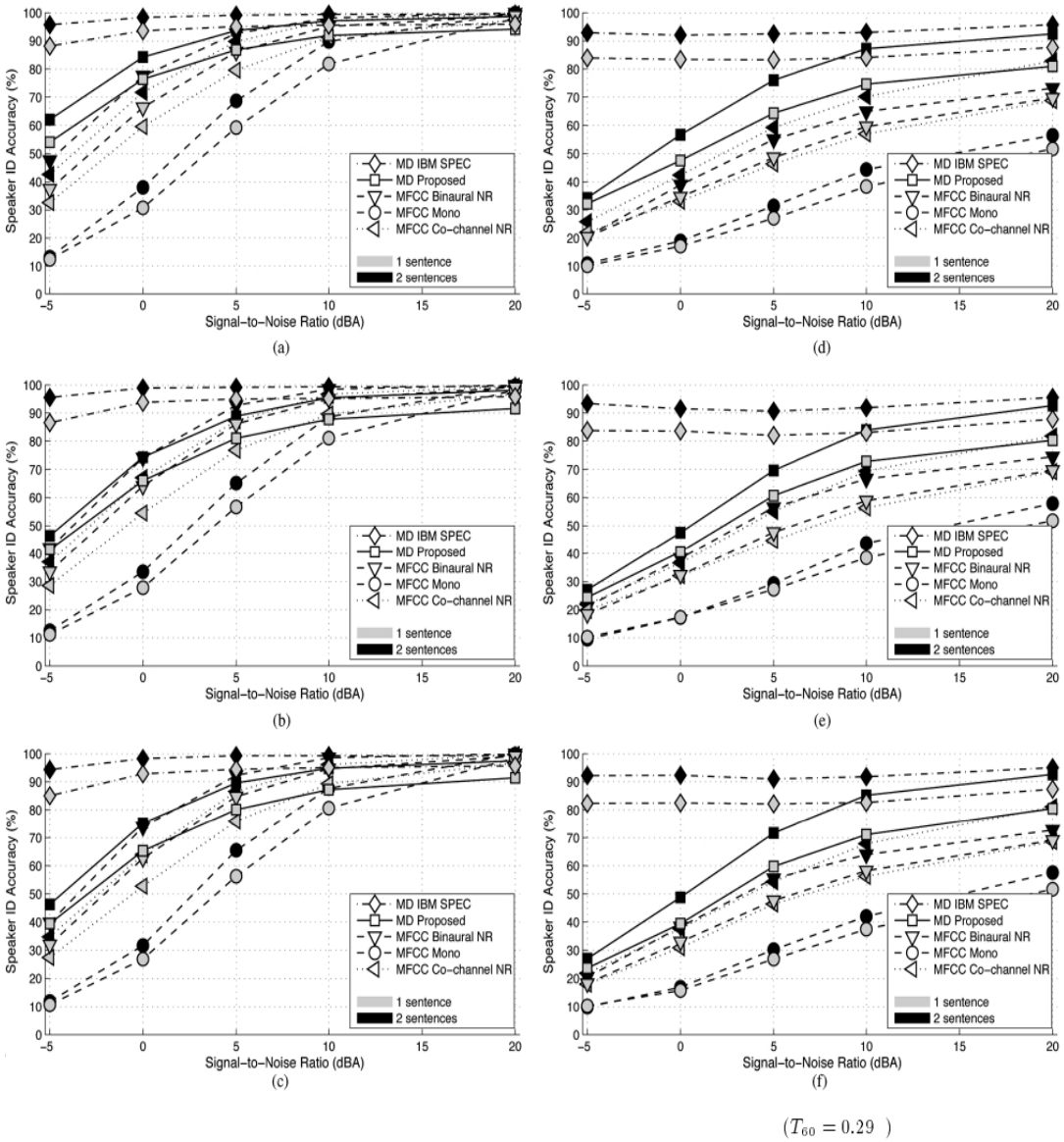


Fig. 4. Experiment 3: Average speaker recognition performance in % for a set of ten speakers in the presence of (A) one, (B) two, and (C) three simultaneously interfering factory noise sources. The average recognition performance is plotted over a series of 20 simulations. Results are presented for three categories of methods, namely the IBM-based MD recognizers (dash-dotted lines), the proposed MD system (solid lines) and the MFCC-based recognizers (dashed lines). The standard error of recognition performance across all 20 simulations was below 3% for all experimental conditions.

sources, most remarkable when comparing results for scenarios with one and two interfering noise sources. This dependency might be related to the fact that the spatial separation between the target and the interfering sources effectively decreases with increasing number of noise sources. The average azimuth spacing for mixtures with 1, 2, and 3 interfering noise sources is 70.4 , 52.3 , and 41.1 , respectively. It is reasonable to consider that the mask computation is more challenging for mixtures with more closely spaced sound sources. The reality is that no such dependency is observed for MD systems that utilize the ideal binary mask suggests that the mask estimation of the proposed method can potentially be improved, especially for multi-source scenarios with closely spaced sound sources.



($T_{60} = 0.29$)

Fig. 5. Experiment 4: Average speaker identification accuracy in % for a set of 34 speakers in reverberant conditions ($T_{60} = 0.29\ s$) in the presence of (A,D) one, (B,E) two, and (C,F) three interfering factory noise sources. Panels (A)-(C) and (D)-(F) show performance for one and two competing target speakers, respectively.

The gray and black symbols decode recognition performance based on one and two sentences, respectively. Results are presented for four categories of methods, namely the IBM-based MD recognizer (dash-dotted lines), the proposed MD system (solid lines), the MFCC-based recognizers (dashed lines) and the MFCC-based Co-channel recognizer (dotted lines).

At higher SNRs (20 dBA), MFCC-based recognizers provide some benefits over the proposed MD system, that is presumably caused by the fact that the distribution of MFCC features is more adequately modeled by Gaussian mixtures with diagonal covariance matrices compared to spectral features. This observation is consistent with results shown previously [22], [52].

To investigate the influence of the speech detection module on speaker identification accuracy, performance is also shown for the proposed system *MD Proposed a priori* which is employing *a priori* knowledge about the azimuth locations of the speech and noise sources. The distance between this method and *MD Proposed* can be interpreted as the error that is introduced by the speech detection module. As we can see in Fig. 4, the performance of both methods is very similar, suggesting that the speech detection module is able to robustly detect the azimuth location of the target. This interpretation is also supported by the low localization error for mixtures with one target speaker, which is reported in Table I. Best results are obtained by the help of MD classifier which is using the ideal binary mask based on the spectral criterion *MD IBM SPEC*, because it considers both the masking effect of interfering noise and the deteriorating effect due to reverberation.

4.4 Experiment 4: Multitalker Speaker Identification

This fourth experiment compares the proposed method with the best performing baseline systems according to the third experiment using acoustic mixtures with one and two concurrently active target speakers. Furthermore, the MFCC-based co-channel recognizer [47] is evaluated. we used a full set of 34 speakers for this experiment. The average speaker identification accuracy is shown in Fig. 5 as a function of the SNR. Results are individually shown for mixtures with one and two target speakers [see panels (A)-(C) and (D)-(F)]. The MFCC-based recognizer *MFCC Binaural NR* which combines better-ear selection and noise reduction is working quite robust for mixtures with one target source and is by far superior to the conventional monaural MFCC-based recognizer. However, this advantage is noticeably reduced when two target speakers are concurrently present. Because the front-end for MFCC feature extraction does not differentiate between target and interfering sources, the resulting feature vector shows some extent properties of all acoustic sources that are available in the acoustic scene. Apparently, the availability of a second target speaker, which is not avoided by the MFCC-based recognizer, creates a systematic bias that clearly limits speaker recognition performance.

The system *MFCC Co-channel NR*, that combines the multitalker training with the noise reduction front-end explained in Section III-B, is able to alleviate the mismatch between the trained speaker models and the observed co-channel mixtures, thus substantially outperforming the monaural MFCC-based recognizers in two-talker mixtures. Note that a considerably larger amount of data is required for training the co-channel system. However, it cannot reach the performance level of the proposed missing data recognizer, which aims at separating the contribution of both speakers.

Generally, the proposed system *MD Proposed* shows a significant performance gain over all MFCC-based recognizers. For mixtures with one target speaker, this benefit is mostly found for very low SNRs. When two target speakers are simultaneously present, however, the advantage of the proposed method covers a great range of SNRs and is especially pronounced at higher SNRs (starting at 5°dBA). This coincides with the SNR at which the speech detection module is still able to predict the azimuth of two speakers within 5° accuracy (see Table I); thus, the binary masks are calculated for the azimuth directions which correspond to the real positions of the speakers.

According to the expectation, the highest speaker recognition accuracy is obtained by the MD classifier *MD IBM SPEC* which is based on *a priori* information about the reliable T-F units. Especially at lower SNRs, there is a substantial difference between the ideal and the proposed MD recognizer, suggesting that there is quite some room for increasing the mask computation. We also investigated the effect of using two sentences to identify the speaker identities. For mixtures with one target speaker, using two sentences for recognition consistently im- proves performance for all methods. However, for mixtures with two concurrently active target speakers, a smaller improvement is found for the MFCC-based systems when two sentences are combined. As previously mentioned, the presence of a second target speaker is likely to cause a mismatch between training and testing, which can obviously not be reduced by increasing the observation time of the classifier. In contrast, the performance gain of *MD Proposed* can be as large as 15% at an SNR of 20°dBA when two sentences are used for recognition.

[3]Note that the co-channel approach is a monaural system. We tested several modifications of the co-channel system and selected the one with the best performance. It is conceivable that the co-channel approach would also benefit from binaural information.

Because of the MD recognizer previously operates on a limited set of T-F units that is believed to consists of reliable information about the target source only, a longer test sequence presumably supplies additional evidence about the speaker identity. In addition, the co-channel system *MFCC Co-channel NR* shows a substantial performance improvement in the range of up to 15% when increasing the time interval used for recognition, most likely due to the reduced mismatch between the training and testing condition.

Lastly, when comparing the total speaker identification accuracy for the set of ten speakers (third experiment) with the full set of 34 speakers (fourth experiment), We can see that the advantage of the MD recognizer over MFCC-based systems decrease as the set of speakers increases. A similar trend was reported in respect of speech recognition [52].

4.5 Experiment 5: Joint Localization and Speaker Recognition

The previous experiment computed the joint localization and speaker identification performance of the proposed method for multi-source mixtures consisting of three interfering factory noise sources, one or two simultaneously active target speakers, and reverberation
($T_{60} = 0.29\ s$).

Similar to the previous experiment, the full set of 34 speakers is used and two sentences are concatenated for recognition.

In order to simultaneously compare the localization accuracy of the target speakers and the recognition accuracy of their identities, the performance of both tasks is jointly visualized by means of a two-by-two

confusion matrix. The localization accuracy shows the percentage of correctly localized target speakers for which the computed azimuth is within an absolute error margin of 5° compared to their real position. The confusion matrices based on the SNR are shown in Fig. 6 for mixtures with one and two target speakers. For each confusion matrix, the sum along the first column shows the speaker identification accuracy, whereas the sum along the first row shows the localization accuracy. The joint localization and recognition performance, which signifies that both tasks were successfully accomplished by the proposed system is represented by the first element of the main diagonal.

We can see that the joint performance is very close to the overall speaker identification accuracy, which denotes that most of the errors are induced by the speaker recognition stage. Indeed, the localization performance of the proposed model is very robust for a wide range of SNRs. Even for mixtures with two simultaneous target speakers and three interfering noise sources, in 96.8% of the cases the azimuth of both speakers is correctly localized for SNRs as low as 5° dBA. However, there is a substantial discrepancy between the localization accuracy and the speaker identification accuracy, which is larger for mixtures with two competing speakers and generally increases with decreasing SNR. This difference may indicate that although the correct azimuth location of the target speaker is available, the accuracy of the computed binary mask at very low SNRs is not enough to robustly determine the identities of the detected speakers. Further investigation is required to improve the quality of the calculated binary mask for complex multi-source scenarios.

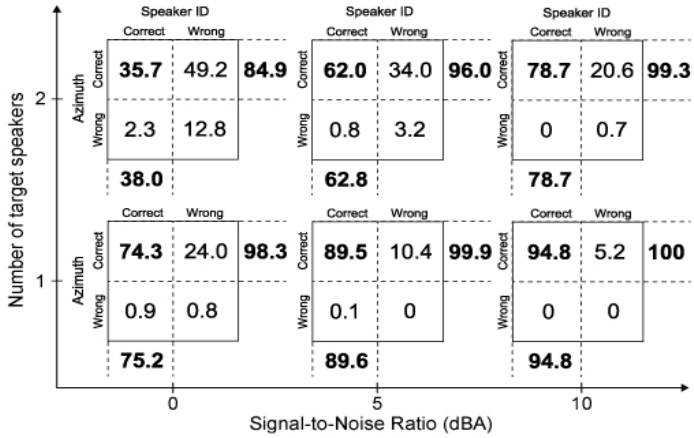


Fig. 6. Experiment 5: SNR-dependent confusion matrices showing the joint localization accuracy and speaker recognition performance of the proposed bin- aural scene analyzer for mixtures consisting of one or two simultaneously active target speakers in the presence of three factory noise sources in a reverberant environment ($T_{60} = 0.29\ s$)

V. DISCUSSION AND CONCLUSION

In this study, we have presented a binaural scene analyzer which is able to jointly localize, detect, and recognize a known number of target speakers in the presence of reverberation and interfering noise. This proposed system contains three main components which are: a binaural front-end for robust localization, a module for speech source detection and a stage for speaker identity recognition. It was shown that this proposed speech detection module is able to robustly detect a known number of target speakers in multi-source scenarios. Depend on this established link between the localization and the recognition stage, this proposed system is able to selectively focus on processing speech sources in the presence of interfering noise. The system does not require *a priori* knowledge about the azimuth position of the target sources, which is often a restriction for practical applications such as hearing aids.

To compare the quality of the computed binary mask of the proposed binaural front-end with the system proposed by Palomäki *et al.* [14], a detailed ROC analysis was performed. This analysis revealed that this proposed system produces binary masks that are closer to the ideal binary masks for both reverberant and anechoic conditions. The proposed front-end has two major benefits: it is designed to operate in reverberant, multi-source scenarios and it jointly analyzes both ITD and ILD cues.

The computed azimuth position of the target speaker can be used to substantially improve performance of MFCC-based recognizers by selecting the *better ear* feature space for recognition. In relation to acoustic scenes with one target speaker,

The improvement in terms of speaker identification accuracy was found to be in the range of 20% at low SNRs. In addition to, moderate improvement was obtained by applying a noise reduction scheme prior to extracting the MFCC coefficients. However, MFCC-based systems only perform well in acoustic scenes with source. This limitation is induced by the front-end for MFCC feature extraction which is not able to distinguish

between target and interfering sources (e.g., the interfering noise sources and concurrent speakers). Whereas MFCC coefficients are, to some extent, able to cope with interfering noise, the presence of a second target speaker clearly biases the resulting MFCC feature vector which consequently limits speaker identification performance. This sensitivity of MFCC-based recognizers to the presence of multiple target speakers can be significantly reduced by the co-channel approach [47], which incorporates a multi-conditional training stage with two-talker mixtures to alleviate the mismatch between training and testing.

Finally, the proposed binaural scene analyzer is more robust than MFCC-based systems, specifically at lower SNRs. Considering acoustic mixtures with a single interfering noise source, the performance of the binaural scene analyzer is close to the classifier that uses the ideal binary mask based on the *a priori* SNR. However, when increasing the number of interfering noise sources, the benefit of the proposed system reduces in comparison to the MFCC-based recognizers. In reality, with decreasing spatial separation between target and interfering sources, it is more difficult to identify reliable T-F units of the target speaker by only exploiting binaural cues. For further improve the computation of the binary mask, the analysis of binaural cues could be extended by additionally exploiting monaural cues such as pitch [17], [19].

Our experimental results shows that there is a significant difference in speaker identification performance when comparing acoustic scenes with one and two simultaneously active target speakers. In the present work, multisource scenarios contains a number of simultaneously active speech and noise sources that were completely overlapping. However, the amount of overlapping speech in a meeting or telephone conversation has been computed to be in the range of 10% [53].The experimental results acheived in this paper are based on simulations also, and further tests with real recordings are required. Future work will focus on more realistic multi-source scenarios with natural overlap and turn-taking.

Furthermore, it was shown that the ideal binary mask, which considers both the effect of reverberation and interfering noise, outperformed the mask based on the *a priori* SNR. This suggests that two mechanisms may be required in order to further improve on existing mask computation techniques: one that segregates the target from the background, and a second one that selects reliable T-F units that are not contaminated by reverberation. A task for future work is to research how to combine the segregation mask, as proposed in this study based on binaural cues, with a mask that assesses the reliability of individual T-F units in terms of reverberation, whether based on modulation analysis [22], [54] or by exploring the effect of temporal masking [55]. In this paper, we assumed prior knowledge about the number of active target speakers that are present in the acoustic scene. An important aspect for future research is to automatically find out the number of active speech sources.

Finally, it was demonstrated that this proposed binaural scene analyzer is able to jointly localize and recognize two simultaneously active target speakers in the presence of three interfering noise sources and reverberation .

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975-979, Sep. 1953.

[2] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acustica*, vol. 86, pp. 117-128, 2000.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[4] *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. Brown, Eds. Hoboken, NJ: Wiley, 2006.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267-285, 2001.

[6] D. Pullella, M. Kühne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. ICASSP*, Las Vegas, NV, 2008, pp. 4833-4836.

[7] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 108-121, Jan. 2012.

[8] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007-4018, Dec. 2006.

[9] N. Li and P. C. Loizou, "Factors influencing glimpsing of speech in noise," *J. Acoust. Soc. Amer.*, vol. 122, no. 2, pp. 1165-1172, Aug. 2007.

[10] D. L. Wang, U. Kjems, M. S. Pedersen, and J. B. Boldt, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2336-2347, Apr. 2009.

[11] D. L. Wang, "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, ch. 12, pp. 181-197.

[12] M. L. Hawley and R. Y. Litovsky, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *J. Acoust. Soc. Amer.*, vol. 115, no. 2, pp. 833-843, Feb. 2004.

[13] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Recognition of reverberant speech using full cepstral features and spectral missing data," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 289-292.

[14] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivated by auditory processing," in *Proc.*

ICASSP, Prague, Czech Republic,2011,pp.5072-5075.
[15] B. D. Simpson, D. S. Brungart, N. Iyer, R. H. Gilkey, and J. T. Hamil, "Detection and localization of speech in the presence of competing speech signals," in *Proc. ICAD*, London, U.K., Jun. 2006, pp. 129-133.
[16] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small- room reverberation," *Speech Commun.*, vol. 43, no. 4, pp. 361-378, 2004.
[17] S. Harding, J. Barker, and G. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58-67, Jan. 2006.
[18] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236-2252, Oct. 2003.
[19] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environ- ments," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 4593-4596.
[20] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1-13, Jan. 2011.
[21] J. Woodruff and D. L. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1856-1866, Sep. 2010.
[22] G. Kidd, Jr., T. L. Arbogast, C. R. Mason, and F. J. Gallun, "The advantage of knowing where to listen," *J. Acoust. Soc. Amer.*, vol. 188, no. 6, pp. 3804-3815, Dec. 2005.
[23] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Commun.*, vol. 43, no. 1-2, pp. 123-142, Jun. 2004.
[24] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1-2, pp. 103-138, Aug. 1990.
[25] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297-336, Oct. 1994.
[26] R. M. Stern, A. S. Zeiberg, and C. Trahiotis, "Lateralization of complex binaural stimuli: A weighted-image model," *J. Acoust. Soc. Amer.*, vol. 84, no. 1, pp. 156-165, Jul. 1988.
[27] T. M. Shackleton, R. Meddis, and M. J. Hewitt, "Across frequency integration in a model of lateralization," *J. Acoust. Soc. Amer.*, vol. 91, no. 4, pp. 2276-2279, Apr. 1992.
[28] T. May, S. van de Par, and A. Kohlrausch, "Binaural detection of speech sources in complex acoustic scenes," in *Proc. WASPAA*, New Paltz, NY, Oct. 2011, pp. 241-244.
[29] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 101-116, Sep. 2005.
[30] M. El-Maliki and A. Drygajlo, "Missing features detection and handling for robust speaker verification," in *Proc. Eurospeech*, Budapest, Hungary, Sep. 1999, pp. 975-978.
[31] M. Cooke and T.-W. Lee, "Speech separation and recognition competition," 2006, accessed on 12th Oct. 2010, [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparation Challenge.htm
[32] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speaker recognition," Speech Research Unit, Defence Research Agency, Malvern, U.K., 1992, Tech. Rep..
[33] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory.*, vol. 28, no. 2, pp. 129-137, Mar. 1982.
[34] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, no. 1, pp. 1-38, 1977.
[35] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 199, no. 3, pp. 1562-1573, Mar. 2006.
[36] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19-41, 2000
[37] B. G. Shinn-Cunningham, J. Schickler, N. Kopˇ and R. Litovsky, co, "Spatial unmasking of nearby speech sources in a simulated anechoic environment," *J. Acoust. Soc. Amer.*, vol. 110, no. 2, pp. 1118-1129, Aug. 2001.
[38] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab, Perceptual Computing, Tech. Rep. #280, 1994.
[39] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, Apr. 1979.
[40] S. M. Schimmel, M. F. Müller, and N. Dillier, "A fast and accurate "shoebox" room acoustics simulator," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 241-244.
[41] *American National Standard Specification for Sound Level Meters*, ANSI/ASA S1.4-1983 (R2001), Amer. Nat. Stand. Inst., 1983.
[42] D. P. W. Ellis, PLP and RASTA (and MFCC, and Inversion) in Matlab, 2005, accessed on 11th October 2011 [Online]. Available:http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat
[43] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 6, pp. 871-879, Jun. 1988.
[44] J. P. Openshaw and J. S. Mason, "On the limitations of cepstral features in noise," in *Proc. ICASSP*, Adelaide, South Australia, Australia, Apr. 1994, pp. 49-52.
[45] S. Tibrewala and H. Hermansky, "Multi-band and adaptation approaches to robust speech recognition," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997, pp. 2619-2622.
[46] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. ICASSP*, Detroit, MI, 1995, vol. 1, pp. 153-156.

[47] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443-445, Apr. 1985.
[48] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M. G. Christensen, S. H. Jensen, and P. Fränti, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proc. ICPR*, Istanbul, Turkey, Aug. 2010, pp. 4565-4568.
[49] T. Fawcett, "An introduction to ROC analysis," *Pattern Recog. Lett.*, vol. 27, no. 8, pp. 861-874, Jun. 2006.
[50] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486-1494, Sep. 2009.
[51] A. Ihlefeld and B. G. Shinn-Cunningham, "Effect of source location and listener location on ILD cues in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, pp. 2598, 2004.
[52] B. G. Shinn-Cunningham, N. Kopˇ and T. J. Martin, "Localizing co, nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, vol. 117, no. 5, pp. 3100-3115, May 2005.
[53] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486-1501, Nov. 2006.
[54] Dr. Joseph Picone, *Fundamentals of speech recognition,. a short course Institute for Signal and Information Processing.* Department of Electrical and Computer Engineering., Mississippi State University
[55] Monson H Hayes, *Digital Signal Processin,* Text book., Schaum's outline.

Proceedings Papers:
[56] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 2001, pp. 1359-1362.